Computational Systems Biology: Biology X

Bud Mishra

Room 1002, 715 Broadway, Courant Institute, NYU, New York, USA

L#5:(Mar-21-2010)
Genome Wide Association Studies

Outline

1. Mendel's Laws
   - Experiments on Garden Peas
   - Statistical Significance
2. Genetic Data
   - Linkage Disequilibrium (LD)

The law of causality ... is a relic of a bygone age, surviving, like the monarchy, only because it is erroneously supposed to do no harm ...

–Bertrand Russell, *On the Notion of Cause*. Proceedings of the Aristotelian Society 13: 1-26, 1913.

Outline

1. Mendel’s Laws
   - Experiments on Garden Peas
   - Statistical Significance
2. Genetic Data
   - Linkage Disequilibrium (LD)

Mendel

- Experimental hybridization in plants – Empirical Studies
- Inheritance of physical units (later dubbed "genes")
- **Principle of Inheritance**: A universal theory to explain how traits in offspring can be predicted from traits in parents.

Mendel’s Analysis

- Rules for
  (a) Predicting genotypes of the offspring from the genotypes of the parents;
  (b) Modeling how genotypes are related to phenotypes.
- **Note**: Genes and genotypes could not be observed (underlying biology of cell-division, fertilization, genomic programs, etc. were unknown to Mendel).
- Mendel proposed *A Genetic Model*: probability distribution for the trait conditional on the underlying genotypes at the hypothetical disease locus.

Mendel’s Models

- **Dichotomous (Mendelian) traits**, **Deterministic Outcomes**, **No co-dependences**, etc.
- These models can be generalized further: using statistical hypothesis testing and with estimation of parameters in the genetic models (Model Selection).

Mendel, 1865, "Experiments in Plant Hybridization"

- Eight years of experiments with garden peas:
- Experiment Design
  1. Very simple (Mendelian) traits
  2. Large number of sample of crosses
  3. Avoiding unintended cross-fertilization
  4. Choosing hybrids with no reduction of fertility (no selection bias)

Relations

- **Traits** ↦ determined by **genotypes** ↦ determined by **Genes**
- Traits are Mendelian; Genotype-Phenotype relation is deterministic (full-penetrance); Genotypes are simple genetic loci (underlying a single a trait); Traits are neutral(!); Genes are bi-allelic (a: wild-type; A: mutant)

Simplicity of Mendelian Experiment

- **Constant Differentiating Characteristics.** Chose simple dichotomous traits and avoided “transitional & blended traits.”
- Used F_1 (first generation hybrid) and P (pure forms) to infer genotypes. Used a self-pollinating (highly inbred) plant which can also cross-pollinate.
- **Underlying Assumptions**: (i) Two genetic variants A and a; (ii) Diploidy: Pure forms are homozygous AA and aa (from self-pollinating inbreds); & (iii) Cross-pollination to create hybrid forms (F_1, F_2, etc.) with heterozygous genotypes Aa.

- Since only one of the two possible phenotypic forms is observed in F_1 hybrids, it's possible to infer the novel association between genotypes and traits.

$$\begin{aligned}\text{Traits appearing in } F_1\ Aa &= \text{Dominant}\\ \text{Traits disappearing from } F_1\ Aa &= \text{Recessive}\end{aligned}$$

equivalently,

$$\begin{aligned}P(\text{recessive form of trait} \mid aa) &= 1\\ P(\text{recessive form of trait} \mid AA) &= 0\end{aligned}$$

$$\begin{aligned}P(\text{dominant form of trait} \mid AA) &= 1\\ P(\text{dominant form of trait} \mid aa) &= 0\end{aligned}$$

Deterministic Model

$$\begin{aligned} \{Aa, AA\} &\mapsto \textbf{DominantTrait} \\ \{aa\} &\mapsto \textbf{RecessiveTrait} \end{aligned}$$

- *A* causes “Dominant Trait.”

Reappearance of the recessive form

- Second generation hybrids F_2 = Offsprings of F_1 hybrids; F_2 has both recessive and dominant forms in the ratio 1 : 3

$$\begin{aligned} P(\text{dominant form of trait} \mid Aa) &= 1 \\ P(\text{recessive form of trait} \mid Aa) &= 0 \end{aligned}$$

- Reappearance of the recessive form – genes for the recessive form remained intact in F_1

$$AA \mapsto 1/4; Aa \mapsto 1/2; aa \mapsto 1/4$$

$$\text{Dominant} \mapsto 3/4; \text{Recessive} \mapsto 1/4$$

$$\text{Dominant} : \text{Recessive} = 3 : 1$$

Mendel’s First Law: Segregation

- One allele of each parent is randomly and independently selected with probability $1/2$ for transmission to the offspring; the alleles unite randomly to form the offspring’s genotype.

Mendel's Second Law: Segregation

- The allele's underlying two or more different traits are transmitted to offspring independent of each other; the transmission of each unit separately follows the first law of segregation.

Genetic Model

- Find the relationship (usually probabilistic) between an individual's genotype and phenotype.
- **Genetic Epidemiology**: Binary trait Y (Affected: $Y = 1$ vs. unaffected: $Y = 0$)

$$Y = g(X_1, X_2, \ldots),$$

where X_i's are (quantitative) *intermediate phenotypes* or *endophenotypes* — **reproducible assessment of the disease features**.

Individual's Phenotype

- Individual's phenotype at a marker = Combination of their two alleles at that locus.
- G is biallelic (also, called di-alleleic)... $A =$ rare/minor/mutant allele; $a =$ frequent/major/wild-type/normal allele
- **Genotypes:** AA & aa are minor & major homozygous, resp. Aa is heterozygous.
- They are sought at DSL: Disease Susceptibility locus.

D-Allele

- *D*-allele is the *Disease variant* or Disease susceptibility allele.
- The genotype-phenotype relation is deterministic or probabilistic.
- The probabilistic relation is described a penetrance function $P(Y|G)$.

$$P(Y=1|G)+P(Y=0|G)=1.$$

- If

$$H_0 : P(Y|G=dd)=P(Y|G=DD)=P(Y|G=Dd)=0,$$

then the disease susceptibility locus G has no effect on the disease status Y.

Mode of Inheritance

- How parameters of the distribution of Y depend on the number of disease allele?
- Four modes of inheritance:
  - **Dominant**
  - **Recessive**
  - **Additive**
  - **Codominant**

$$P(Y = 1|DD) = 1; P(Y = 1|dd) = 0;$$

$$P(Y = 1|Dd) = \alpha P(Y = 1|DD) + P(Y = 1|dd).$$

- Dominant $\mapsto \alpha = 1$; Additive $\mapsto \alpha = 1/2$; Recessive $\mapsto \alpha = 0$
- Co-dominant makes no assumption about the relation among the three penetrance function.

Reduced Penetrance Models

- For some $0 \leq \beta_0 < 1$ & $0 < \beta_1 \leq 1$,

$$P(Y = 1|DD) = \beta_1; P(Y = 1|dd) = \beta_0.$$

- **Phenocopies**: Disease could also be caused by another genetic locus or possibly a non-genetic co-variate.

GLM: Generalized Linear Model

- $g(\cdot)$ is a *link function*; $E(\cdot) =$ expectation:

  $$g(E(Y|X)) = \beta_0 + X'\beta_1.$$

  $X =$ Coding of genotype in terms of the mode of inheritance.

- **Logistic link**

  $$\log \frac{E(Y|X)}{1 - E(Y|X)} = \beta_0 + X'\beta_1.$$

- **Log(relative risk link**

  $$\log E(Y|X) = \beta_0 + X'\beta_1.$$

- Null Hypothesis

  $$H_0 : \beta_1 = 0.$$

  Acceptance implies no relationship between the gene and the trait.

Difficulties

- Effects leading to spurious causal explanations:
- **Confounding and effect mediation**
- A *confounder* is a variable that is: (1) associated with the exposure (cause) variable; (2) independently associated with the outcome (effect) variable; and (3) not in the causal pathway between exposure and disease.
- *Example: Heavy alcohol consumption (the exposure) is associated with the total cholesterol level (the outcome). However smoking tends to be associated with heavy alcohol consumption. Smoking is also associated with high cholesterol levels among the individuals who are not heavy alcohol users.*
- A confounder is defined as a clinical or demographic variable that is associated with the genotype and the trait under investigation.

Difficulties

- A variable lying on the causal pathway between the predictor and the outcome is called an *effect mediator* or causal pathway variable.
- Genotype affects the trait through alteration of the mediator variable.
- A particular SNP variant may make an individual more likely to smoke and smoking would then cause cancer. Here smoking is an effect mediator.

Contingency Table

- Three genotypes for a given SNP: *homozygous wildtype* *aa*, *heterozygous* *Aa* and *homozygous rare/* *AA*.
- The data can be represented by the 2 × 3 contingency table. See below.
- **Odds Ratio**: *Ratio of the odds of disease among the exposed to the odds of disease among the unexposed.*
- Genotype $\equiv$ *exposure*

	Gen: *aa*	Gen: *Aa*	Gen: *AA*	
Dis: $+$	n_{11}	n_{12}	n_{13}	$n_{1\cdot}$
Dis: $-$	n_{21}	n_{22}	n_{23}	$n_{2\cdot}$
	$n_{\cdot 1}$	$n_{\cdot 2}$	$n_{\cdot 3}$	n

Odds Ratio

- Odds Ratio:

$$OR = \frac{Pr(D^+|E^+)/[1 - Pr(D^+|E^+)]}{Pr(D^+|E^-)/[1 - Pr(D^+|E^-)]}$$

- In genetics, we calculate the *OR* for each genotype with relation to the homozygous wildtype genotype, *AA*.

$$OR_{aa,AA} = \frac{(n_{11}/n_{.1})/(n_{21}/n_{.1})}{(n_{13}/n_{.3})/(n_{23}/n_{.3})} = \frac{n_{11}n_{23}}{n_{21}n_{13}}$$

Dichotomized Contingency Table

- Dichotomizing genotype priors
- $E^+ = \{Aa, aa\}$ and $E^- = \{AA\}$
- The data can be represented by the 2×2 contingency table. See below.

	Gen: $\{aa, Aa\}$	Gen: AA	
Dis: $+$	n_{11}	n_{12}	$n_{1\cdot}$
Dis: $-$	n_{21}	n_{22}	$n_{2\cdot}$
	$n_{\cdot 1}$	$n_{\cdot 2}$	n

Odds Ratio

- Odds Ratio:

$$\widehat{OR} = \frac{(n_{11}/n_{\cdot 1})/(n_{21}/n_{\cdot 1})}{(n_{12}/n_{\cdot 2})/(n_{22}/n_{\cdot 2})} = \frac{n_{11}n_{22}}{n_{21}n_{12}}$$

Fisher’s Exact Test

- What is the probability of getting the 2×2 table by *chance*

$$p = \binom{n_{1\cdot}}{n_{11}} \binom{n_{2\cdot}}{n_{21}} \Big/ \binom{n}{n_{\cdot 1}} = \frac{n_{1\cdot}! n_{2\cdot}! n_{\cdot 1}! n_{\cdot 2}!}{n! n_{11}! n_{12}! n_{21}! n_{22}!}$$

- This formula gives the exact probability of observing this particular arrangement of the data, assuming the given marginal totals, on the null hypothesis that te two categories of genotypes are equally likely to have the disease.
- In other words, the probability p indicates how well the data fit the hypothesis: “the single or double mutation ($A \mapsto a$) cause the disease.”
- If $p \ll \theta$ (i.e., the probability is very very small), we can reject the null hypothesis, and conclude that “the mutation ($A \mapsto a$) has a necessary causal role in the disease.”

Fisher's exact test

- Fisher's exact test is a statistical test used to determine if there are nonrandom associations between two categorical variables. — E.g., Genotypes and a Categorical Trait.
- Let there exist two such variables X and Y, with m and n observed states, respectively.
- Now form an $m \times n$ matrix in which the entries a_{ij} represent the number of observations in which $x = i$ and $y = j$. Calculate the row and column sums R_i and C_j, respectively, and the total sum

$$N \sum_i R_i = \sum_j C_j.$$

of the matrix.

- Then calculate the conditional probability of getting the actual matrix given the particular row and column sums, given by

$$P_{cutoff} = \frac{(R_1!R_2!\cdots R_m!)(C_1!C_2!\cdots C_n!)}{N!\prod_{ij} a_{ij}!}$$

  which is a multivariate generalization of the **hypergeometric probability function**.

- Now find all possible matrices of nonnegative integers consistent with the row and column sums R_i and C_j. For each one, calculate the associated conditional probability using this formula, where the sum of these probabilities must be 1.

- To compute the P-value of the test, the tables must then be ordered by some criterion that measures dependence, and those tables that represent equal or greater deviation from independence than the observed table are the ones whose probabilities are added together.
- There are a variety of criteria that can be used to measure dependence. In the 2×2 case, which is the one Fisher looked at when he developed the exact test, either the Pearson chi-square or the difference in proportions (which are equivalent) is typically used.
- Other measures of association, such as the likelihood-ratio-test, -squared, or any of the other measures typically used for association in contingency tables, can also be used.

Pearson's chi-square (χ^2) test

- Null hypothesis states that the "*frequency distribution of certain events observed in a sample is consistent with a particular theoretical distribution*."
- The events considered must be mutually exclusive and have total probability 1. The events each cover an outcome of a categorical variable.
- Used for (1) Tests of goodness of fit and (2) Tests of independence.
- **Example**: Test the hypothesis that an ordinary six-sided die is "fair," i.e., all six outcomes are equally likely to occur.

- A test of independence assesses whether paired observations on two variables, expressed in a contingency table, are independent of each other. E.g., association between a categorical exposure (genotype) and categorical disease variable (trait).
- In case of a 2×2 contingency table test of no association between rows and columns $\equiv$ the single null hypothesis $H_0 : OR = 1$. That is, expected count

$$n_{11} \approx n \cdot Pr(D^+) Pr(E^+) = n(n_{1.}/n)(n_{.1}/n) = E_{11} = n_{1.} n_{.1}/n.$$

General Scheme

- The expected count for the (i,j) cell is given by $E_{ij} = n_{i.}n_{.j}/n$, where $i = 1, \cdots, r$ (rows) and $j = 1, \cdots, c$ (columns).
- Let the corresponding observed cell counts be denoted by O_{ij}.
- Pearson's χ^2-statistics is given by

$$\chi^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \sim \chi^2_{(r-1)(c-1)}.$$

- That is, this statistics has χ^2-distributions with $(r-1)(c-1)$ degrees of freedom.

p-value

- The χ^2 statistic can then be used to calculate a *p*-value by comparing the value of the statistic to a χ^2-distribution.
- A χ^2 probability $\leq$ 0.05 is commonly interpreted as justification for rejecting the null hypothesis that the row variable is unrelated (that is, only randomly related) to the column variable.
- Fisher’s exact test is preferable when at least 20% of the expected cell counts are small ($E_{ij} < 5$).

Cochran-Armitage (C-A) Trend Test

- The Cochran-Armitage test for trend is typically used in categorical data analysis when some categories are ordered.
- For instance, with a biallelic locus with three genotypes $aa = 0$, $aA = 1$, and $AA = 2$, ordered by the number of A alleles, it can be used to test for association in a 2×3 contingency table.

- Define a statistic

$$T = \sum_{i=1}^{3} t_i (n_{1i} n_{2.} - n_{2i} n_{1.}),$$

  where t_i's are weights.
- Null hypothesis (H_0) of no association indicates that

$$E(T) = 0, var(T) = \frac{(n_{1.} n_{2.})}{n} \sum_{i=1}^{3} t_i^2 n_{.i}(n - n_{.i}) - 2 \sum_{i=1}^{2} \sum_{j=i+1}^{3} t_i t_j n_{.i} n_{.j}$$

  p-values are computed assuming that $T/\sqrt{var(T)} \sim N(0,1)$.

	Gen: *aa*	Gen: *Aa*	Gen: *AA*	
Dis: $+$	n_{11}	n_{12}	n_{13}	$n_{1\cdot}$
Dis: $-$	n_{21}	n_{22}	n_{23}	$n_{2\cdot}$
	$n_{\cdot 1}$	$n_{\cdot 2}$	$n_{\cdot 3}$	n

Another Interpretation

- If we let p_j be the probability of the disease for the jth genotype column, and S_j is the score for the jth column, i.e. $S_j =$ number of A alleles $+1$, then the C-A test is testing for the trend by solving the following linear regression
$$p_j = \alpha + \beta S_j.$$
- The null hypothesis H_0 is then tested y checking the trend: $\beta = E(T) = 0$.

Correlation

- The *correlation coefficient* between two random variables is defined as the ratio of the covariance between these two variables and the product of their standard deviations.

$$cc(X, Y) = \frac{cov(X, Y)}{\sqrt{var(X) var(Y)}}.$$

- The correlation coefficient is a measure of linear association between two variables and takes values between -1 and $+1$.
- Two most common sample-based estimates of the correlation coefficient: (1) Pearson's product-moment correlation coefficient and (2) Spearman's rank correlation coefficient. Pearson's coefficient is highly sensitive to outliers!

Outline

1. Mendel's Laws
   - Experiments on Garden Peas
   - Statistical Significance
2. Genetic Data
   - Linkage Disequilibrium (LD)

LD and HWE

- Two important concepts to explore
- (1) **LD**: Linkage Disequilibrium
- (2) **HWE**: Hardy-Weinberg Equilibrium
- *Concepts regarding the genetic component of the data — No connection to traits.*

LD and HWE

- Both LD and HWE are measures of allelic association....
- LD measures the associations among sites along the genome
- HWE measures the association at a single site between pair of homologous chromosmes.

Linkage Disequilibrium (LD)

- Association between two adjacent variant sites become lost over time as recombination events occur in the region separating them. Asymptotically the genomes will go to linkage equilibrium, making all the sites acting independently.
- If the sites are all independent then only the “causal variant” site will contribute to the “probability raising” and will not have any “screening off” due to some other confounding (correlated) sites.
- However, in general the variant sites are not yet in linkage equilibrium and there exist strong dependence among the sites.

- **The SNP sites that are usually analyzed in GWAS could be within genes, but may not be functional.** That is, these SNP sites may not directly cause the disease.
- Usually "tag SNPs" that are analyzed are selected to represent the haplotypes occurring within a haplotype block–they are non-functional but closely associated to functional/causal SNPs.
- These sites are likely to be *associated with* disease because they are in LD (**Linkage Disequilibrium**) with the *functional variant*.
- LD is measured in terms of two closely related measures: D' and r^2.
- These measures are very closely related to Pearson's χ^2-statistics.

LD: D'

- Consider the distribution of alleles for n individuals across two sites: Assume that the two sites are *independent of each other* – **in Linkage Equilibrium**.
- *The presence of an allele at one site does not influence the particular allele observed at the second site.*
- Assume: At site 1 the alleles are A and a, with population frequencies p_A and p_a, respectively. At site 2 the alleles are B and b, with population frequencies p_B and p_b, respectively.

	Site 2		
	B	b	
Site A	$n_{11} = Np_Ap_B$	$n_{12} = Np_Ap_b$	$n_{1\cdot} = Np_A$
1 a	$n_{21} = Np_ap_B$	$n_{22} = Np_ap_b$	$n_{2\cdot} = Np_a$
	$n_{\cdot 1} = Np_B$	$n_{\cdot 2} = Np_b$	$N = 2n$

LD: D'

- If sites 1 and 2 are in fact associated with one another, then the observed counts will deviate from the numbers shown in the earlier table.
- Represent the deviation by a single scalar D.
- $H_0 : D = 0$ corresponds to the null hypothesis that the two sites are independent (in LE: Linkage Equilibrium).

	Site 2		
	B	b	
Site A	$n_{11} = N(p_A p_B + D)$	$n_{12} = N(p_A p_b - D)$	$n_{1\cdot}$
1 a	$n_{21} = N(p_a p_B - D)$	$n_{22} = N(p_a p_b + D)$	$n_{2\cdot}$
	$n_{\cdot 1}$	$n_{\cdot 2}$	$N = 2n$

Estimating D'

- D can be expressed in terms of the joint probability of A and B and the product of the individual allele probabilities:

$$D = p_{AB} - p_A p_B.$$

- Note that we can estimate D as

$$\widehat{D} = \widehat{p_{AB}} - \widehat{p_A}\widehat{p_B} = \widehat{p_{AB}} - (n_{1\cdot}/N)(n_{\cdot 1}/N).$$

- $\widehat{p_{AB}}$ has to be estimated by an MLE estimator

Estimating D'

- Let $\theta = (p_{AB}, p_{Ab}, p_{aB}, p_{ab})$ be estimated from the genotype counts from two biallelic loci

$$\begin{aligned}
&\log L(\theta | n_{11}, \ldots, n_{33}) \\
&\quad \propto \quad (2n_{11} + n_{12} + n_{21}) \log p_{AB} + (2n_{13} + n_{12} + n_{23}) \log p_{Ab} \\
&\qquad + (2n_{31} + n_{21} + n_{32}) \log p_{aB} + (2n_{33} + n_{32} + n_{23}) \log p_{ab} \\
&\qquad + n_{22} \log(p_{AB} p_{ab} + p_{Ab} p_{aB}).
\end{aligned}$$

	Site 2		
	BB	*Bb*	*bb*
Site *AA*	n_{11}	n_{12}	n_{13}
1 *Aa*	n_{21}	n_{22}	n_{23}
aa	n_{31}	n_{32}	n_{33}

- One can estimate D by first substituting $p_A p_B + D$ for p_{AB}, $p_A p_b - D$ for p_{Ab}, etc. and solve the maximization problem for $\widehat{D}$ using numerical optimization.
- Alternatively, write $p_{Ab} = p_A - p_{AB}$, $p_{aB} = p_B - p_{AB}$, and $p_{ab} = 1 - p_A - p_B - p_{AB}$, and estimate p_{AB}. Solve for $\widehat{D} = \widehat{p_{AB}} - \widehat{p_A}\widehat{p_B}$.
- A rescaled value of D, given by D' is used for a measure of LD:

  $$D' = \frac{|D|}{D_{\max}},$$

  where $D_{\max}$ bounds D from above:

  $$D_{\max} = \begin{cases} \min(p_A p_b, p_a p_B), & \text{if } D > 0; \\ \min(p_A p_B, p_a p_b), & \text{otherwise.} \end{cases}$$

- Note that $0 \leq D' \leq 1$
- If D' is close to 1, then the two sites are assumed to be in "complete LD." The sites are in the same haplotype block.
- If D' is close to 0, then the two sites are assumed to be independent — with a recombination hot-spot separating them. The sites belong to two distinct adjacent haplotype blocks.

The quantity r^2

- The quantity r^2, measuring LD, is based on Pearson's χ^2-statistic for the test of no association.
- Consider an $r \times c$ contingency table corresponding to the counts of individuals with two bi-allelic sites: Site 1: A, a and site 2: B, b.

  $$\chi_1^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}},$$

  where $i = 1, \ldots, r$; $j = 1, \ldots, c$; and O_{ij} and E_{ij} are respective observed and expected cell counts for the i, jth cell of an $r \times c$ table.
- r^2 is defined as

  $$r^2 = \chi_1^2 / N.$$

Relation between D' and r^2

$$(O_{ij} - E_{ij})^2 = (ND)^2.$$

Thus

$$\begin{aligned}
\chi_1^2 &= \sum_{ij} \frac{(ND)^2}{E_{ij}} \\
&= (ND)^2 \left(\frac{1}{Np_A p_B} + \frac{1}{Np_A p_b} + \frac{1}{Np_a p_B} + \frac{1}{Np_a p_b} \right) \\
&= ND^2 \left(\frac{p_a p_b + p_a p_B + p_A p_b + p_A p_B}{p_A p_B p_a p_b} \right) \\
&= \frac{ND^2}{p_A p_B p_a p_b}
\end{aligned}$$

Relation between D' and r^2

- In summary,

$$r^2 = \chi_1^2/N = \frac{D^2}{p_A p_B p_a p_b}.$$

- Thus r^2 is simply D^2, further adjusted by the marginal probabilities.
- r^2 is usually preferred, because of its straightforward relationship with the χ^2 statistics and the null hypothesis H_0 that the two sites are independent.

Caveat

- With the currently available technology (e.g., genotype sequencing), haplotypes are not observed — so the cell counts in the contingency tables are inferred.
- The estimation process (MLE or EM), introduces further errors into the r^2 measures — making it highly unreliable.
- Additionally, Pearson's χ^2-test assumes independent observations — which may be violated in the absence of HWE (Hardy-Weinberg Equilibrium). Note that the contingency table includes two observations per person

Summary

- D' and r^2 are both *measures* of linkage disequilibrium between loci; they estimate the amount of association between two sites.
- Conclusions from these must be drawn with caution — as they depend on certain implicit assumptions that are often violated.
- The Key problem: **Haplotype Phasing Problem**

LD Blocks

- Determine whether a group of adjacent loci are in LD.
- A measure of LD across a region (comprising multiple SNPs) is the average of all pairwise measures of D'

$$\bar{D}' = \frac{1}{n_L} \sum_{i,j \in L} D'_{ij},$$

where

- L is a set of loci within a region of interest
- D'_{ij} is the measure of LD between loci i and j for $i, j \in L$
- n_L is the number of ways of choosing two loci from the set L (i.e., $\binom{|L|}{2}$)
- the summation is over all such pairs of loci

LD Blocks

- Through characterization of regions of high average LDs, a genome (i.e., human's) can be partitioned in to *LD Blocks*.
- These blocks are separated by (recombination) *hotspots* – regions in which recombination events might have occurred with very high frequencies (and likely to happen in the future).
- In general, *alleles tend to be more correlated within an LD block* than across...

SNP tagging

- Once regions of high LD are identified, we will aim to determine the smallest subset of SNPs that characterizes the variability in the region — this process is called *SNP tagging* and the selected SNPs are called **Tag SNPs**.
- Example: Consider two SNPs (i and j) that are in perfect LD so that $D'_{i,j} = 1$. Genotyping both SNPs are unnecessary as their relationship is *deterministic* — knowledge of the genotype of one SNP completely defines the genotype of the second and there is no need to sequence both loci.
- Few (say 3 - 5) well-defined tag SNPs capture a substantial majority of the genetic variability within an LD block.

TAG SNPs

- Note: in general, tag SNPs are correlated with the true disease causing variant – but are not typically functional themselves.
- LD blocks differ substantially across race and ethnicity groups: It's shorter in Black/non-Hispanics than White and Hispanics.
- African population has much more genetic variability. It is older with many more recombination events than the European population.
- *A tag SNP may capture information on the true disease-causing variant in one racial group, but not another.*
- Thus in any GWAS, understanding population substructures and its effect on measures of LD is CRUCIAL!!!!

LD and Population Stratification

- **Population Stratification**: Presence of multiple subgroups (of sub-populations) among which there is minimal mating and gene-flow.
- *Ignoring population stratification in a sample could lead to confounding conclusions.*
- Population ad mixtures pose additional problems.
- Simpson's paradox – Yule-Simpson Effect. This paradox occurs in the presence of a confounding variable that is not properly accounted for in the analysis.

Hardy-Weinberg Equilibrium (HWE)

- HWE denotes independence of alleles at a single site between two homologous chromosomes.
- For instance, consider the simple case of biallelic SNP with genotypes *AA*, *Aa* and *aa*.
- HWE implies that the probability of an allele occurring on one homolog does not affect which allele will be present on the second homolog:

$$p_{AA} = p_A^2, p_{Aa} = p_{aA} = p_A p_a, \text{ and } p_{aa} = p_a^2,$$

where

$$p_A + p_a = 1.$$

Violation of HWE

- Tests of HWE include Pearson's χ^2-test and Fisher's exact test.
- When more than 20% of the expected counts are less than five, Fisher's exact test is recommended. The χ^2-test is computationally efficient but relies on asymptotic theories.
- The test s are based on the 2×2 table of genotypes at a single locus, as shown below:

	Homolog 2		
	A	a	
Homolog A	n_{11}	n_{12}	$n_{1\cdot}$
1 a	n_{21}	n_{22}	$n_{2\cdot}$
	$n_{\cdot 1}$	$n_{\cdot 2}$	n

χ^2-test

- Note: n_{11} and n_{22} are the counts for major and minor homozygous individuals: AA and aa, respectively.
- The two heterozygous genotypes are indistinguishable: One can only observe $n^*_{12} = n_{12} + n_{21}$.
- The expected values, corresponding to observations $O_{11} = n_{11}$, $O_{12} = n^*_{12}$ and $O_{22} = n_{22}$ are

$$E_{11} = np_A^2, E_{12} = 2np_A(1-p_A), \text{ and } E_{22} = n(1-p_A)^2.$$

- The χ^2-statistic:

$$\chi^2 = \sum_{i=1,2} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \sim \chi^2_1.$$

HWE

- Let i index the individual in a sample of n independent individuals from a population with allele frequency p_A. Let $X_i (i = 1, \ldots, n)$ denote the number of A alleles the ith person in the sample. Let X_+ denote the summation of X_i over all n individuals.

$$X_i \sim Bin(2, p_A); E(X_i) = 2p_A; \mathrm{var}(X_i) = 2p_A q_A.$$

- An estimate for p_A is

$$\widehat{p_A} = \frac{1}{2n} \sum_i X_i = \frac{2n_{11} + n_{12}^*}{2n}.$$

- Thus

$$E(\widehat{p_A}) = \frac{1}{2n} \sum_i E(X_i) = p_A, \quad \& \quad \mathrm{var}(\widehat{p_A}) = \frac{1}{4n^2} \sum_i \mathrm{var}(X_i) = \frac{p_A q_A}{2n}.$$

HWE

- In large populations, $\widehat{p_A}$ is approximately $N(p_A, p_A q_A/2n)$.
- To test the null hypothesis $H_0 : p_A = p_0$ at the α-level, we reject if the magnitude of

$$Z = \frac{\sqrt{2n}(\widehat{p_A} - p_0)}{\sqrt{p_0(1-p_0)}}$$

is greater than the $(1-\alpha)/2$-percentile $(Z_{(1-\alpha)/2})$ of a standard normal distribution.

HWD

- A statisticaly significant test of HWE suggests that *the SNP under investigation is in* **Hardy-Weinberg Disequilibrium** *(HWD).*
- HWD is usually assumed to be resulting from self-seleting mates: *non-random mating*.
- Deviation from HWE may also indicate non-neutral evolution: *positive or negative selection*
- Question: What is the relationship between HWE and population substructure?

HWE and Population Substructure

- HWE is based on the assumptions of: random mating, no inbreeding, infinite population size, discrete generations, equal allele frequencies in males nd females, and no mutation, migration or selection.
- Note:
  1. HWE implies constant allele frequencies over generations.
  2. HWE is violated in the presence of population admixtures.
  3. HWE is violated in the presence of population stratification.
- These observations and the corresponding statistical tests allow one to understand the population substructures and use them to correct the causal analysis of GWAS.

Allele Frequencies over Generations

- The genotype of a parent (at a single biallelic locus):

$$pr(AA) = p_A^2, pr(Aa) = 2p_A q_A, \text{ and } pr(aa) = q_A^2,$$

where $q_A = p_a = (1 - p_A)$.

- The inheritance pattern. The conditional probability that an offspring inherits allele y, given that the parent has genotype X is $pr(y|X)$.

$$pr(A|AA) = 1, pr(A|Aa) = 1/2, \text{ and } pr(A|aa) = 0.$$

- Thus the population frequency of the allele A in the next generation is given by

$$\begin{aligned} pr(A) &= pr(A|AA)pr(AA) + pr(A|Aa)pr(Aa) + pr(A|aa)pr(aa) \\ &= p_A^2 + p_A q_A + 0 = p_A. \end{aligned}$$

Population Admixtures

- Population Admixtures occur as a result of matings between two populations for which alele frequencies differ.
- Assume that the two populations have two different frequencies for the allele A: p_{1A} and p_{2A}.
- Then the offsprings resulting from random matings of the two populations (assuming infinite populations sizes) will have frequencies:

$$\begin{aligned} pr(AA) &= p_{1A}p_{2A}, pr(Aa) = p_{1A}q_{2A} + p_{2A}q_{1A}, \text{ and} \\ pr(aa) &= q_{1A}q_{2A}. \end{aligned}$$

Note: $q_{iA} = 1 - p_{iA}, i = 1, 2$.

Population Stratification

- Population stratification is the combination of populations in which breeding occurs within but not between sub-populations.
- Within each sub-populations, we may have HWE (since the observed counts are as expected under random mating).
- Assume population 1 has allele frequency: $pr(A) = p_{1A}$ and population 2: $pr(A) = p_{2A}$. Assume that the two populations are of equal size, but $p_{1A} \ll p_{2A}$. The combined frequency is $p_A = (p_{1A} + p_{2A})/2$, but

$$\begin{aligned} pr(AA) &= (p_{1A}^2 + p_{2A}^2)/2 \approx p_{2A}^2/2, \text{ but} \\ p_A^2 &= (p_{1A} + p_{2A})^2/4 \approx p_{2A}^2/4. \end{aligned}$$

Failure of HWE

- Counting:

$$\begin{aligned} Pr(X=0) &= p_{aa}, \\ Pr(X=1) &= p_{Aa}, \\ Pr(X=2) &= p_{AA}. \end{aligned}$$

- Thus

$$E(X) = 0 \cdot p_{aa} + 1 \cdot p_{Aa} + 2 \cdot p_{AA} = 2p_A.$$

&

$$\mathrm{var}(X) = 0 \cdot p_{aa} + 1 \cdot p_{Aa} + 4 \cdot p_{AA} - E(X)^2 = p_{Aa} = 2p_A q_A.$$

Failure of HWE: Population Stratification

- Assume a population with K strata, with allele frequencies p_k and strata frequencies s_k, for $k = 1, \ldots, K$.
- By definition the allele frequencies in the total population is $p = \sum_k s_k p_k$.
- Now

$$\begin{aligned}
Pr(X=0) &= \sum_k s_k q_k^2 = 1 - 2p + E(p_k^2) - p^2 + p^2 \\
&= q^2 + var(p_k) \\
Pr(X=1) &= 2\sum_k s_k p_k q_k = 2p - 2E(p_k^2) + 2p^2 - 2p^2 \\
&= 2pq - 2var(p_k), \\
Pr(X=2) &= \sum_k s_k p_k^2 = E(p_k^2) - p^2 + p^2 = p^2 + var(p_k).
\end{aligned}$$

Failure of HWE: Population Stratification

- Thus

$$\begin{aligned} E(X) &= 0\cdot(q^2+var(p_k))+1\cdot(2pq-2var(p_k)) \\ &\quad +2\cdot(p^2+var(p_k)) \\ &= 2p^2+2pq=2p. \end{aligned}$$

&

$$\begin{aligned} var(X) &= 0\cdot(q^2+var(p_k)+1\cdot(2pq-2var(p_k)) \\ &\quad +4\cdot(p^2+var(p_k))-E(X)^2 \\ &= 2pq+2var(p_k). \end{aligned}$$

- With a stratified population, $var(X)$ is inflated and the frequency of heterozygosity is reduced.

Failure of HWE: Population Inbreeding

- With inbreeding, there is a positive probability that an individual inherits the exact same A (or a) allele from both parents – increasing homozygosity...
- F is defined to be the *inbreeding coefficient* $=$ is the probability that a randomy sampled individual will inherit the same copy from both parents.

$$\begin{aligned} Pr(X = 0) &= Fq_A + (1 - F)q_A^2, \\ Pr(X = 1) &= 2p_Aq_A(1 - F), \\ Pr(X = 2) &= Fp_A + (1 - F)p_A^2. \end{aligned}$$

Failure of HWE

- Thus

$$
\begin{aligned}
E(X) &= 2[Fp_A + (1-F)p_A^2] + 2p_A q_A (1-F) = 2p_A \\
var(X) &= 4[Fp_A + (1-F)p_A^2] + 2p_A q_A (1-F) - 4p_A^2 \\
&= 2p_A q_A (1+F).
\end{aligned}
$$

- There is a deficit of heterozygotes relative to HWE ... *Loss of Heterozygosity (LOH)*.
- Further $var(X)$ is inflated.

Hardy's Law

- Mendelian genetics: it was not then known how it could cause continuous characteristics. Udny Yule (1902) argued against Mendelism because he thought that dominant alleles would increase in the population.
- The American William E. Castle (1903) showed that without selection, the genotype frequencies would remain stable. Karl Pearson (1903) found one equilibrium position with values of $p = q = 1/2$.
- Reginald Punnett introduced the problem to G. H. Hardy, a British mathematician... who found biologists' use of mathematics as “very simple.”
- The principle was known as Hardy's law in the English-speaking world until 1943, when Curt Stern pointed out that it had first been formulated independently in 1908 by the German physician Wilhelm Weinberg.

Hardy's Letter

- “To the Editor of Science:
- “I am reluctant to intrude in a discussion concerning matters of which I have no expert knowledge, and I should have expected the very simple point which I wish to make to have been familiar to biologists. However, some remarks of Mr. Udny Yule, to which Mr. R. C. Punnett has called my attention, suggest that it may still be worth making...
- “Suppose that *Aa* is a pair of Mendelian characters, *A* being dominant, and that in any given generation the number of pure dominants (*AA*), heterozygotes (*Aa*), and pure recessives (*aa*) are as $p : 2q : r$.

Hardy's Letter

- "Finally, suppose that the numbers are fairly large, so that mating may be regarded as random, that the sexes are evenly distributed among the three varieties, and that all are equally fertile. A little mathematics of the multiplication-table type is enough to show that in the next generation the numbers will be as $(p+q)^2 : 2(p+q)(q+r) : (q+r)^2$, or as $p_1 : 2q_1 : r_1$, say.
- "The interesting question is — in what circumstances will this distribution be the same as that in the generation before? It is easy to see that the condition for this is $q^2 = pr$. And since $q_1^2 = p_1 r_1$, whatever the values of p, q, and r may be, the distribution will in any case continue unchanged after the second generation."

Heterozygote advantage

- HWE may be violated under selection pressure: E.g., *heterozygote advantage*, or *heterotic balancing selection*. ... An individual who is heterozygous at a particular gene locus has a greater fitness than a homozygous individual.
- Example: Sickle cell anemia... a hereditary disease that damages red blood cells.
- Sickle cell anemia is caused by the inheritance of a variant hemoglobin gene (HgbS) from both parents. In these individuals hemoglobin (protein in red blood cells that carries oxygen to the tissues) is extremely sensitive to oxygen deprivation causing short life expectancy.

Heterozygote advantage

- However, a person who inherits the sickle cell gene from one parent and a normal hemoglobin gene (HgbA) from the other parent (a carrier of the sickle cell trait) has a normal life expectancy. The heterozygote is resistant to the malarial parasite which kills a large number of people each year.

Heterozygote advantage

- HgbS, which in the homozygous state causes sickle-cell anemia, is distributed throughout sub-Saharan Africa, the Mediterranean, the Middle East, and parts of India; the frequency of the carrier state ranges from 5 to over 40 percent.
- HgbE, the most common structural hemoglobin in the world population, is confined to the eastern regions of the Indian subcontinent, Myanmar, and Southeast Asia. Its frequency varies; carrier rates of over 60 percent of the population occur in eastern Thailand and parts of Cambodia.

[End of Lecture #5]